Università degli Studi di Padova
Dipartimento di Scienze Statistiche

Corso di Laurea Triennale in
Statistica per le Tecnologie e le Scienze

Analisi del dataset `Mussels`

Relazione di:
Stefano Terrone
Leonardo Perin

Anno accademico:
2024/2025

0.1 Introduzione

Il dataset `mussels`*[1]*, contiene le misurazioni di varie specie di molluschi bivalvi in 44 fiumi facenti parte della East Coast degli Stati Uniti d'America.
Il termine "mussels" è un termine generico che indica varie famiglie di molluschi bivalvi, sia di acqua dolce che di acqua salata, che hanno in comune un guscio allungato e asimmetrico*[2]*. In particolare, nello studio sono state analizzate 79 specie di molluschi, appartenenti alla sottoclasse dei *Palaeoheterodonta* e alla famiglia delle *Unionidae*, detti anche cozze di acqua dolce. I fiumi osservati fanno parte della East Coast degli USA e comprendono i fiumi compresi tra la baia di St. Lawrence (situata in Canada) e lo stato Alabama. Dunque, sono presenti sia fiumi che sfociano nell'Atlantico orientale, sia i fiumi che sfociano nel Golfo del Messico. Lo scopo di questo studio è di valutare quali fattori influiscano sul numero di specie di cozze in un fiume. Per fare ciò sono state osservate altre 9 variabili come il numero totale di specie di cozze presenti in un fiume, l'area totale del fiume, la distanza da fiumi principali e anche fattori di qualità delle acque. La variabile `Species`, è il numero di specie di cozze presenti nel fiume osservato. La variabile `Area` è la misurazione in *miglia quadrate* dell'area di drenaggio di un fiume. L'area è stata misurata usando un planimetro polare su una mappa sulle risorse idriche fornita dal dipartimento sulla geologia degli USA *(USGS)*. La variabile `ln(Area)` è il logaritmo naturale dell'area del fiume. Per valutare la dispersione delle specie di cozze nell'area geografica di riferimento, si è deciso di designare alcuni fiumi come fiumi di origine delle varie specie di cozze. I fiumi scelti sono: il sistema fluviale dell'Alabama-Coose *(AC)*, il fiume Apalachicola *(AP)*, il fiume Savannah *(SV)*, e il fiume St. Lawrence *(SL)*. I primi due fiumi (AP e AC) sfociano nel Golfo del Messico, mentre gli altri due (SV e SL) sfociano nell'Atlantico orientale. Altri criteri che hanno deciso la scelta dei fiumi sono: importanza geografica, numero di specie di cozze presenti e la loro dispersione nei fiumi limitrofi. Data la irregolarità del corso di un fiume, al fine di misurare la distanza dei fiumi osservati dai 4 fiumi principali, non è stata usata una misura lineare della distanza, ma si è deciso di misurare la distanza in termini di prossimità. Viene così sfruttata la caratteristica dei fiumi di scorrere in modo lineare lungo la costa. Ogni fiume viene quindi visto come il passo di un saltello, e la distanza viene stimata semplicemente contando quanti fiumi separano il fiume osservato da quelli principali. Le variabili che misurano la distanza sono 4: `Stepping stones to AC`, ovvero i "saltelli" necessari a raggiungere il sistema fluviale Alabama-Coosa; `Stepping stones to AP`, "saltelli" dal fiume Apalachicola; `Stepping stones to SL`, "saltelli" dalla foce del fiume St. Lawrence e, infine, `Stepping stones to SV`, "saltelli" dalla foce del fiume Savannah. Per valutare la qualità delle acque sono state misurate tre variabili: `Nitrate`, che rappresenta la concentrazione in *parti per milione* di nitrato NO_3*[3]* nelle acque del fiume. Questo ione è fondamentale per la produzione di alghe e batteri che rappresentano la principale fonte di cibo per le cozze*[4]*; `Solid Residue`, che rappresenta il residuo fisso nelle acque, ovvero la quantità in *parti per milione* di sali minerali e oligoelementi presenti in soluzione o sospensione nell'acqua del fiume. Questa variabile è stata misurata facendo prima evaporare l'acqua a 180° C per una ora e, successivamente, pesando ciò che rimane. Risulta utile perchè permette di valutare la concentrazione di alcuni sali come il sodio, il potassio, il calcio e il magnesio e la salinità delle acque; `Hydronium`, la concentrazione in *grammi-ioni per* 10^7 di idronio (H_3O^+) nell'acqua *[6]*. Questa quantità è ottenuta tramite l'operazione inversa del calcolo del pH e rappresenta l'antilogaritmo in base 10 del suo negativo. Questa misurazione torna utile in quanto

il pH, e in particolare la presenza dello ione idrogeno (H^+), è un importante fattore nella calcificazione del guscio delle cozze*[5]*. Non viene misurato direttamente lo ione idrogeno (detto anche idrone) perchè non può esistere in soluzione acquosa allo stato libero.
Le analisi sono state eseguite con il software R nella versione 4.2.3.
(https://www.r-project.org/).
Il livello di significatività è fissato al 5%.

1. Analisi esplorative

1.1 Analisi Univariate

Il database è composto da 44 osservazioni, una per ciascun fiume, e 9 varabili. Le varabili rappresentano varie caratteristiche dei fiumi come il numero di specie di cozze presenti, la dimensione del fiume, la distanza da alcuni fiumi importanti, e vari indicatori sulla qualità delle acque.

	Min.	1° qt.	Mediana(IQR)	Media(sd)	3° qt.	Max
`Species`	2.00	8.00	10.00(5.0)	11.25(5.99)	13.00	33.00
`Area`	349	2115	4315(7797.5)	6590(6016)	9912	27900
`Stepping stones to AC`	1.00	7.00	15.50(15.25)	15.34(9.19)	22.25	33.00
`Stepping stones to AP`	0.00	4.00	12.00(14.25)	12.02(8.24)	18.25	28.00
`Stepping stones to SV`	0.00	5.00	7.00(13.25)	8.136(8.94)	11.00	21.00
`Stepping stones to SL`	4.00	16.75	22.00(0.775)	22.16(1.84)	30.00	36.00
`Nitrate`	0.100	0.600	0.800(0.775)	1.495(1.84)	1.375	8.700
`Solid Residue`	29.0	56.5	78.0(64.0)	112.4(17.3)	120.5	520.0
`Hydronium`	0.200	1.00	1.600(2.2)	3.631(6.03)	3.200	32.00
`ln(Area)`	5.855	7.657	8.370(1.54)	8.331(1.07)	9.201	10.24

Tabella 1: Tabella di sintesi

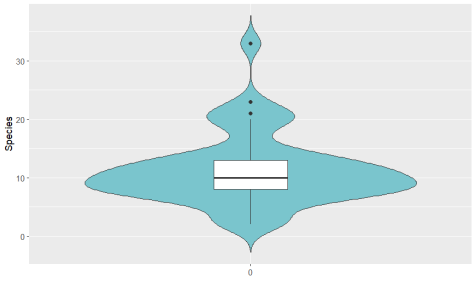

Figura 1: Violinplot della variabile `Species`

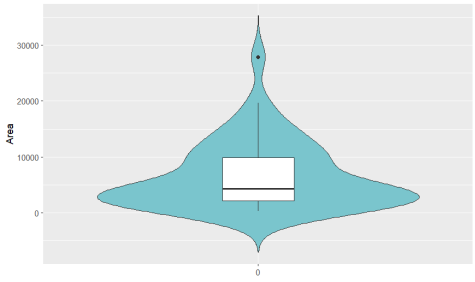

Figura 2: Violinplot della variabile `Area`

La variabile `Species`, rappresentante il numero di specie di cozze nel fiume, ha una distribuzione asimmetrica a destra come evidenziato dal violinplot. Sono presenti 4 outliers: i fiumi Cooper-Santee e Savannah con 21 specie, il fiume Escambia con 23 specie e il fiume Apalachicola con 33 specie *(Figura 1)*. `Species` ha un valore minimo di 2 specie (nel fiume Waccasassa) e un massimo di 33 specie (nel fiume Apalachicola). La mediana è di 10 specie, la media è di 11.25 specie.

La variabile `Area` segue una distribuzione asimmetrica a destra *(Figura 2)*, con un valore minimo di 349 *sq mi (square mile)* e un massimo di 27900 *sq mi*, corrispondente al fiume Susquehanna, che rappresenta l'unico outliers. La mediana è di 4315 *sq mi*, la media è di 6590 *sq mi (Tabella 1)*.

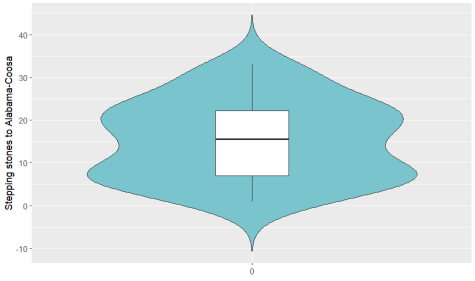

Figura 3: Violinplot della variabile `Stepping stones to AC`

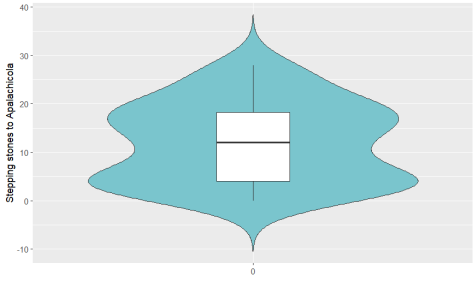

Figura 4: Violinplot della variabile `Stepping stones to AP`

La distanza dal sistema fluviale di Alabama-Coose (`Stepping stones to AC`) segue una distribuzione simmetrica *(Figura 3)*, suggerendo che i fiumi sono equamente distribuiti rispetto a questo sistema fluviale. La distanza minima è di 1 e la massima di 33. La distanza media è di 15.34, e quella mediana di 15.50 "saltelli".

Anche la distanza dal fiume Apalachicola (`Stepping stones to AP`) segue una distribuzione simmetrica *(Figura 4)*, con una distanza minima di 0 (coincidente con il fiume stesso) e una massima di 28. La distanza media è di 12.02, e quella mediana di 12.00 "saltelli" *(Tabella 1)*.

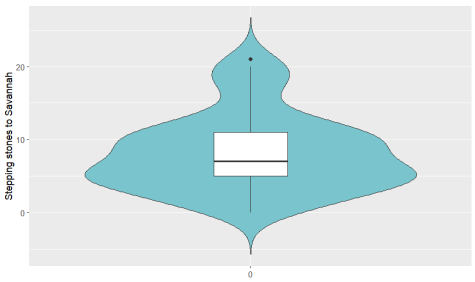

Figura 5: Violinplot della variabile `Stepping stones to SV`

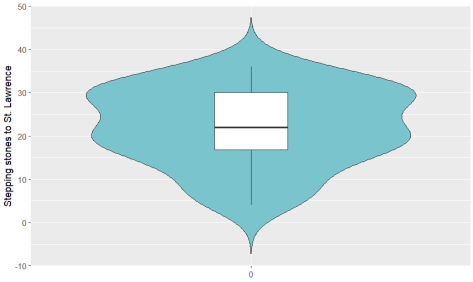

Figura 6: Violinplot della variabile `Stepping stones to SL`

La distanza dal fiume Savannah (`Stepping stones to SV`) ha una distribuzione simmetrica anche se presenta una leggera coda lunga a destra *(Figura 5)*, con una distanza minima di 0 (coincidente con il fiume stesso) e una massima di 21, corrispondente al fiume Penobscot, situato a Nord e outliers. La distanza media è di 8.136, e quella mediana di 7.00. Il valore molto basso della mediana e la presenza di una piccola coda lunga a destra suggeriscono che la maggior parte dei fiumi osservati sono molto vicini al fiume Savannah.

La distanza dal fiume St. Lawrence (`Stepping stones to SL`) ha una distribuzione simmetrica *(Figura 6)*, ma presenta anche una leggera coda lunga a sinistra. La distanza minima di 4 e una massima di 36. La distanza media è di 22.16, e quella mediana di 22.00. Il valore elevato della di media e mediana, in aggiunta alla leggera coda a sinistra, suggerisce che la maggior parte dei fiumi sono distanti al fiume St. Lawrence *(Tabella 1)*.

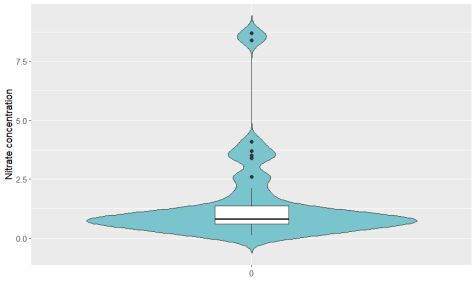

Figura 7: Violinplot della variabile `Nitrate`

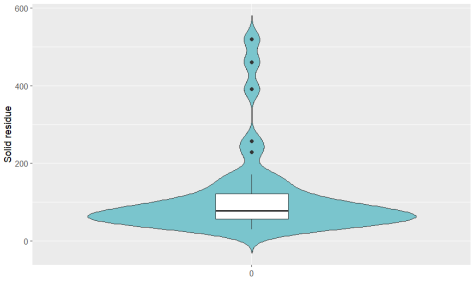

Figura 8: Violinplot della variabile `Solid Residue`

La variabile `Nitrate` ha una distribuzione asimmetrica a destra. Sono presenti 6 outliers: i fiumi Susquehanna, Thames, Potomac, Roanoke, Balckstone e Delaware che presentano tutti una elevata concentrazione di nitrato, rispettivamente di 3.4, 3.5, 3.7, 4.1, 8.4 e 8.7 *(Figura 7)*. Il valore minimo è di 0.100 *ppm (parts per million)* e il massimo di 8.700 *ppm*. La mediana è di 0.800 *ppm*, la media è di 1.495 *ppm*.

La variabile `Solid Residue` ha anche essa una distribuzione asimmetrica a destra. Sono presenti 5 outliers corrispondenti a fiumi con una elevata quantità di residuo fisso nella acque. Essi sono: il Withlacooche (229 *ppm*), il Peace (257 *ppm*), il Waccasassa (461 *ppm*), l'Alafia (520 *ppm*) e il St. Johns, Fla. (391 *ppm*) *(Figura 8)*.

La concentrazione minima è di 29.0 *ppm* e la massima di 520.0 *ppm*. La mediana è di 78.0 *ppm*, la media è di 112.4 *ppm (Tabella 1)*.

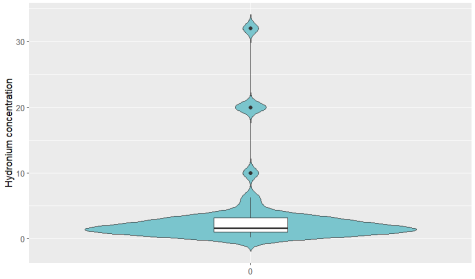

Figura 9: Violinplot della variabile `Hydronium`

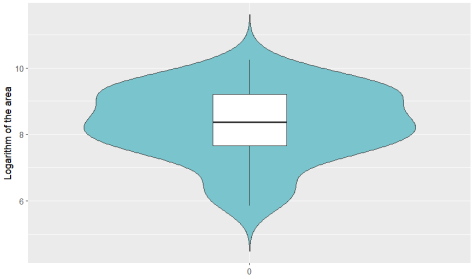

Figura 10: Violinplot della variabile `ln(Area)`

La variabile `Hydronium` ha una distribuzione asimmetrica a destra. Sono presenti 4 outliers, tutti corrispondenti a fiumi con una elevata quantità di idronio. I fiumi sono: il Waccamaw, il St. Marys, il Blackstone e il Satilla *(Figura 9)*. Il valore minimo di idronio è di 0.200, il massimo di 32.00. La mediana è di 1.600, la media è di 3.631.

La variabile `ln(Area)` segue una distribuzione simmetrica *(Figura 10)*, con un valore minimo di 5.855 e un massimo di 10.24. La mediana è di 8.370, la media è di 8.331 *(Tabella 1)*.

1.2 Analisi Bivariate

Si valuta la relazione tra le varie variabili. Delle 36 bivariate totali, 16 bivariate sono risultate significative *(Tabella 2)*.

	A	H^+	SR	NO_3	SL	SV	AP	AC
Sp	1	0	0	0	0	0	0	0
AC	0	1	0	1	1	1	1	
AP	0	1	0	1	1	1		
SV	0	0	0	1	1			
SL	0	1	0	1				
NO_3	0	0	1					
SR	0	1						
H^+	0							

Tabella 2: Tabella che indica la significatività delle correlazioni: 1 indica una bivariata significativa, 0 indica una bivariata non significativa.
In particolare Sp: `Species`, AC: `Stepping stones to AC`, AP: `Stepping stones to AP`, SV: `Stepping stones to SV`, SL: `Stepping stones to SL`, NO_3: `Nitrate`, SR: `Solid Residue`, H^+: `Hydronium`, A: corrispondente sia a `Area` che a `ln(Area)`

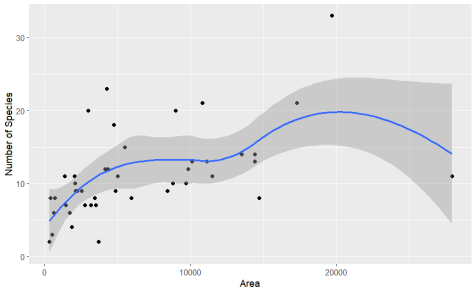

Figura 11: Diagrammma di dispersione tra `Species` e `Area`, con curva LOESS *(LOcally Estimated Scatterplot Smoothing)* in blu e intervallo di confidenza al 95% in grigio

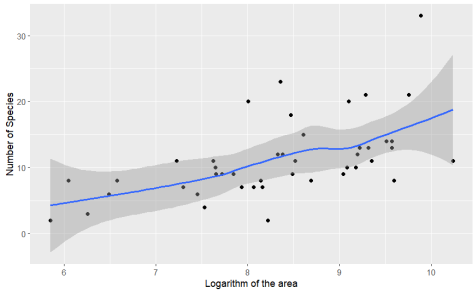

Figura 12: Diagrammma di dispersione tra `Species` e `ln(Area)`, con curva LOESS in blu e intervallo di confidenza al 95% in grigio

Dal diagramma di dispersione tra il numero di specie di cozze e l'area *(Figura 11)* si nota una relazione positiva, ma molto curvilinea tra le due variabili. La correlazione di Spearman è 0.64, indicando una relazione positiva tra le due variabili; dunque all'aumentare dell'area del fiume aumenta anche il numero di specie di cozze presenti. Il test per verificare se la correlazione è significativa (S=5089.2, p-value¡0.001) porta

a rifiutare l'ipotesi che la correlazione non sia significativa che conferma la presenza di una relazione positiva tra le due variabili.

Dal diagramma di dispersione tra il numero di specie di cozze e il logaritmo dell'area *(Figura 12)* si nota una relazione positiva tra le due variabili. La relazione però, appare più lineare rispetto a quella tra l'area e il numero di specie *(Figura 11)*. La correlazione di Spearman è 0.64, uguale a quella osservata in precedenza. Le conclusioni sono le stesse: all'aumentare del logaritmo dell'area del fiume aumenta anche il numero di specie di cozze presenti. Anche il test per verificare se la correlazione è significativa (S=5089.2, p-value¡0.001) risulta uguale al test fatto precedentemente. Dunque, si rifiuta l'ipotesi che la correlazione non sia significativa.

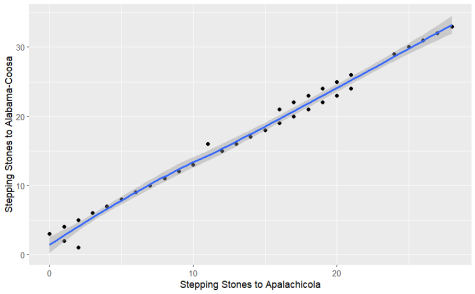

Figura 13: Diagramma di dispersione tra `Stepping stones to AC` e `Stepping stones to AP`, con curva LOESS in blu e intervallo di confidenza al 95% in grigio

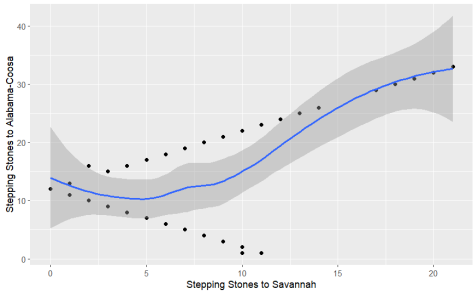

Figura 14: Diagrammma di dispersione tra `Stepping stones to AC` e `Stepping stones to SV`, con curva LOESS in blu e intervallo di confidenza al 95% in grigio

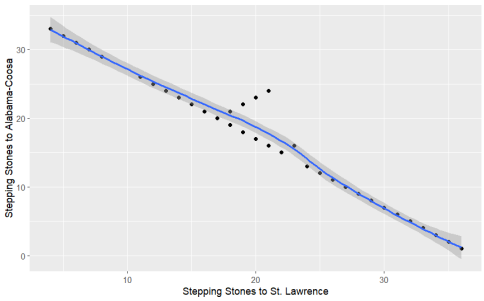

Figura 15: Diagramma di dispersione tra `Stepping stones to AC` e `Stepping stones to SL`, con curva LOESS in blu e intervallo di confidenza al 95% in grigio

Dal diagramma di dispersione tra la distanza dei vari fiumi osservati dal sistema fluviale Alabama-Coose e dal fiume Apalachicola *(Figura 13)* si nota una relazione positiva, quasi lineare, tra le due variabili. La correlazione di Spearman è 0.99, indicando una forte relazione positiva tra le due variabili; dunque all'aumentare della distanza di un fiume dal sistema fluviale Alabama-Coose aumenta anche la distanza dal fiume Apalachicola. Questa relazione è dovuta al fatto che i due fiumi sono molto vicini tra loro, scorrono in parallelo, e sfociano entrambi nel Golfo del Messico, e

quindi la distanza da uno implica la distanza dall'altro. Il test per verificare se la correlazione è significativa (S=89.42, p-value¡0.001) porta a rifiutare l'ipotesi che la correlazione non sia significativa.

Dal diagramma di dispersione tra la distanza dei vari fiumi osservati dal sistema fluviale Alabama-Coose e dal fiume Savannah *(Figura 14)* si nota una relazione tendenzialmente positiva, ma molto curvilinea tra le due variabili. In particolare si nota come nella prima del grafico i punti divergono tra di loro. Questo è dovuto al fatto che il fiume Savannah e il sistema fluviale Alabama-Coose scorrono perpendicolarmente tra di loro, quasi a formare un angolo retto. Quindi una porzione di fiumi (quelli situati a Sud), mentre si allontananano dal fiume Savannah si avvicinano al sistema fluviale Alabama-Coose. Al contrario, i fiumi situati a Nord si allontananano parallamente ad entrambi i fiumi che spiega la crescita positiva nella seconda parte del grafico. La correlazione di Spearman è 0.54, indicando una modesta relazione positiva tra le due variabili. La correlazione è positiva; dunque all'aumentare della distanza di un fiume dal sistema fluviale Alabama-Coose aumenta anche la distanza dal fiume Savannah. Il test per verificare se la correlazione è significativa (S=6483.6, p-value¡0.001) porta a rifiutare l'ipotesi che la correlazione non sia significativa.

Dal diagramma di dispersione tra la distanza dei vari fiumi osservati dal sistema fluviale Alabama-Coose e dal fiume Saint Lawrence *(Figura 15)* si nota una relazione negativa, abbastanza lineare, tra le due variabili. La correlazione di Spearman è -0.97, indicando una forte relazione negativa tra le due variabili; dunque all'aumentare della distanza di un fiume dal sistema fluviale Alabama-Coose diminuisce la distanza dal fiume St. Lawrence. Questo è dovuto dalla posizione dei due fiumi. Essi, infatti, si trovano alle due estremità della regione osservata, con il fiume St. Lawrence situato più a Nord-Est e il sistema fluviale Alabama-Coose situato invece a Sud-Ovest. Quindi allontanandosi dal sistema fluviale Alabama-Coose ci si avvicina al fiume St. Lawrence, e viceversa. Il test per verificare se la correlazione è significativa (S=28075, p-value¡0.001) porta a rifiutare l'ipotesi che la correlazione non sia significativa.

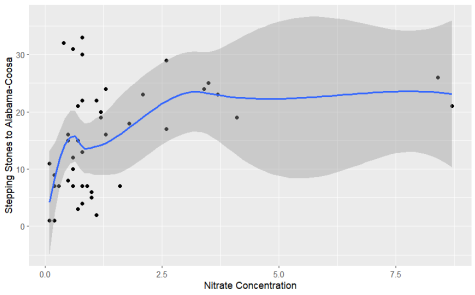

Figura 16: Diagrammma di dispersione tra `Stepping stones to AC` e `Nitrate`, con curva LOESS in blu e intervallo di confidenza al 95% in grigio

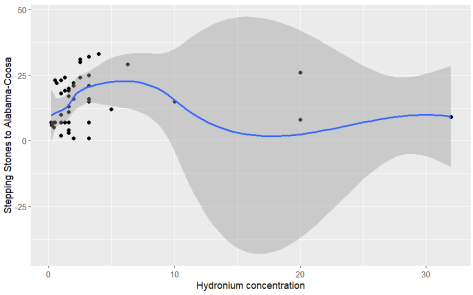

Figura 17: Diagramma di dispersione tra `Stepping stones to AC` e `Hydronium`, con curva LOESS in blu e intervallo di confidenza al 95% in grigio

Dal diagramma di dispersione tra la distanza dei vari fiumi osservati dal sistema fluviale Alabama-Coose e la concentrazione di nitrato nelle acque dei fiumi osservati *(Figura 16)* si osserva una relazione molto curvilinea. In particolare si nota come i fiumi più vicini al sistema fluviale Alabama-Coose presentano una minore concentrazione di nitrato. La concentrazione di nitrato aumenta all'aumentare della distanza

dal sistema fluviale Alabama-Coose, ma solo fino ad un certo punto. Infatti, dopo una certa distanza, la concentrazione di nitrato inizia a diminuire. La correlazione di Spearman è 0.42, indicando una lieve relazione positiva tra le due variabili. Il test per verificare se la correlazione è significativa (S=8173.9, p-value=0.004) porta a rifiutare l'ipotesi che la correlazione non sia significativa.

Dal diagramma di dispersione tra la distanza dei vari fiumi osservati dal sistema fluviale Alabama-Coose e la variabile `Hydronium` *(Figura 17)* si nota come la maggior parte dei fiumi presenti valori bassi di concentrazione di idronio. Fanno eccezione 5 fiumi che invece presentano valori molto alti di concentrazione di idronio. La correlazione di Spearman è 0.33, indicando una lieve relazione positiva tra le due variabili. Il test per verificare se la correlazione è significativa (S=9506.5, p-value=0.029) porta a rifiutare l'ipotesi che la correlazione non sia significativa,.

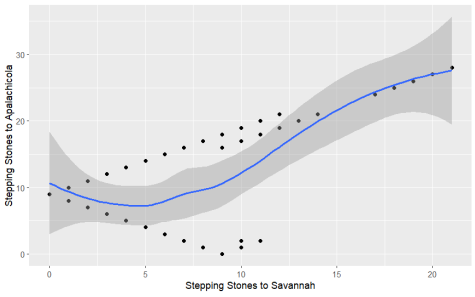

Figura 18: Diagrammma di dispersione tra `Stepping stones to AP` e `Stepping stones to SV`, con curva LOESS in blu e intervallo di confidenza al 95% in grigio

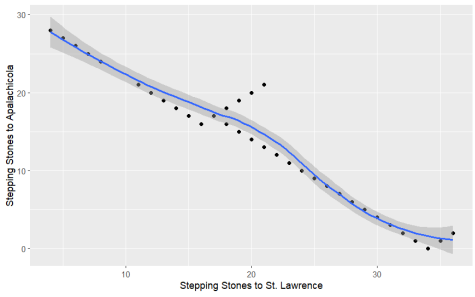

Figura 19: Diagramma di dispersione tra `Stepping stones to AP` e `Stepping stones to SL`, con curva LOESS in blu e intervallo di confidenza al 95% in grigio

Dal diagramma di dispersione tra la distanza dei vari fiumi osservati dal fiume Apalachicola e dal fiume Savannah *(Figura 18)* si nota una relazione tendenzialmente positiva, ma molto curvilinea tra le due variabili. Il grafico mostra una situazione simile a quella osservata tra la distanza dei fiumi dal sistema fluviale Alabama-Coose e dal fiume Savannah in *Figura 14*. Anche in questo caso, infatti, si osserva nella prima parte un divergere delle distanze e successivamente un crescere lineare delle stesse. Ciò è dovuto alla posizione del fiume Apalachicola che è molto vicino al sistema fluviale Alabama-Coose e quindi questo causa un comportamento simile. La correlazione di Spearman è 0.55, indicando una modesta relazione positiva tra le due variabili. Il test per verificare se la correlazione è significativa (S=6436.4, p-value¡0.001) porta a rifiutare l'ipotesi che la correlazione non sia significativa.

Dal diagramma di dispersione tra la distanza dei vari fiumi osservati dal fiume Apalachicola e dal fiume Saint Lawrence *(Figura 19)* si evince una situazione analoga a quella discussa precedentemente. Si nota una relazione negativa, abbastanza lineare, tra le due variabili molto simile a quella in *Figura 15* La correlazione di Spearman è -0.97, indicando una forte relazione negativa tra le due variabili; dunque all'aumentare della distanza di un fiume dal sistema fluviale Apalachicola diminuisce la distanza dal fiume St. Lawrence. Il test per verificare se la correlazione è significativa (S=27892, p-value¡0.001) porta a rifiutare l'ipotesi che la correlazione non sia significativa.

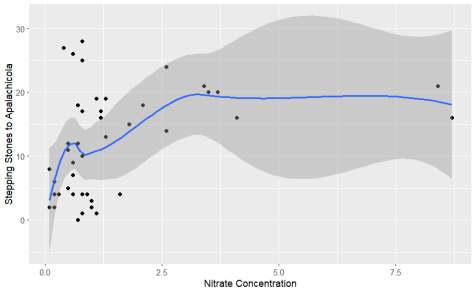

Figura 20: Diagrammma di dispersione tra `Stepping stones to AP` e `Nitrate`, con curva LOESS in blu e intervallo di confidenza al 95% in grigio

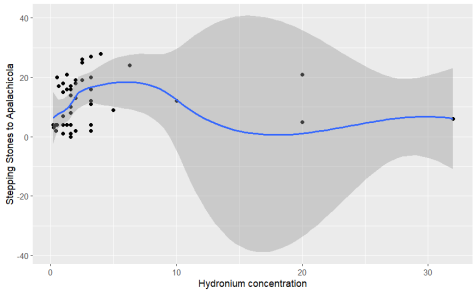

Figura 21: Diagramma di dispersione tra `Stepping stones to AP` e `Hydronium`, con curva LOESS in blu e intervallo di confidenza al 95% in grigio

Dal diagramma di dispersione tra la distanza dei vari fiumi osservati dal fiume Apalachicola e la concentrazione di nitrato nelle acque dei fiumi osservati *(Figura 20)* si nota come i fiumi più vicini al fiume Apalachicola presentano una bassa concentrazione di nitrato, seguita da un aumento di essa mentre ci si allontana dal fiume e infine una diminuzione della concentrazione di nitrato. La correlazione di Spearman è 0.40, indicando una lieve relazione positiva tra le due variabili. Il test per verificare se la correlazione è significativa (S=8454.7, p-value=0.006) porta a rifiutare l'ipotesi che la correlazione non sia significativa.

Dal diagramma di dispersione tra la distanza dei vari fiumi osservati dal fiume Apalachicola e la concentrazione di idronio *(Figura 21)* si nota come la maggior parte dei fiumi sia schiacciata nella prima parte del grafico, con valori bassi di concentrazione di idronio. Solo 5 fiumi, presentano valori molto alti di concentrazione di idronio. La correlazione di Spearman è 0.33, indicando una lieve relazione positiva tra le due variabili. Il test per verificare se la correlazione è significativa (S=9438.8, p-value=0.026) porta a rifiutare l'ipotesi che la correlazione non sia significativa.

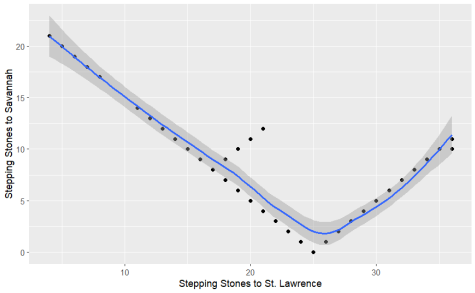

Figura 22: Diagrammma di dispersione tra `Stepping stones to SV` e `Stepping stones to SL`, con curva LOESS in blu e intervallo di confidenza al 95% in grigio

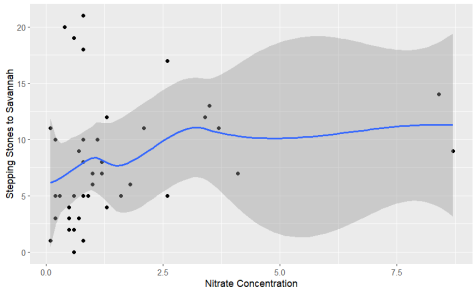

Figura 23: Diagramma di dispersione tra `Stepping stones to SV` e `Nitrate`, con curva LOESS in blu e intervallo di confidenza al 95% in grigio

Dal diagramma di dispersione tra la distanza dei vari fiumi osservati dal fiume Savannah e dal fiume St. Lawrence *(Figura 22)* si nota una relazione inizialmente negativa che, una volta raggiunto il fiume Savannah, cambia drasticamente e cresce

positivamente. Questo andamento è dovuto alla posizione dei due fiumi. Entrambi i fiumi sfociano nell'Atlantico; il fiume St. Lawrence nasce più a Nord, nella regione dei Grandi Laghi e sfocia in Canada, il fiume Savannah invece è situato più a Sud, in Carolina del Sud. La relazione è inizialmente negativa poiché allontanandosi dal fiume St. Lawrence, e procedendo verso Sud, ci si avvicina al fiume Savannah. Una volta raggiunto il fiume Savannah, le distanze tra i due fiumi aumenteranno in parallelo. La correlazione di Spearman è -0.51, indicando una relazione negativa tra le due variabili. Il test per verificare se la correlazione è significativa (S=21406, p-value=0.0004) porta a rifiutare l'ipotesi che la correlazione non sia significativa.

Dal diagramma di dispersione tra la distanza dei vari fiumi osservati dal fiume Savannah e la concentrazione di nitrato nelle acque dei fiumi osservati *(Figura 23)* si nota un andamento simile a quello della *Figura 16* e della *Figura 20*, anche se la curva LOESS appare meno definita. Come nella figure precedenti, i fiumi più vicini al fiume Savannah presentano una bassa concentrazione di nitrato, seguita da un aumento di essa mentre ci si allontana dal fiume e infine una diminuzione della concentrazione di nitrato. La correlazione di Spearman è 0.36, di poco inferiore a quella tra `Stepping stones to AC` e `Nitrate` `Stepping stones to AP` e `Nitrate` rispettivamente di 0.42 e 0.40. La relazione è lieve ed è positiva. Il test per verificare se la correlazione è significativa (S=9025.2, p-value=0.015) porta a rifiutare l'ipotesi che la correlazione non sia significativa.

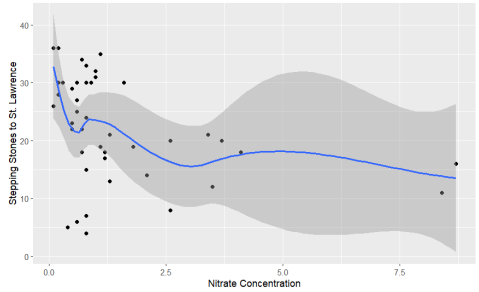

Figura 24: Diagrammma di dispersione tra `Stepping stones to SL` e `Nitrate`, con curva LOESS in blu e intervallo di confidenza al 95% in grigio

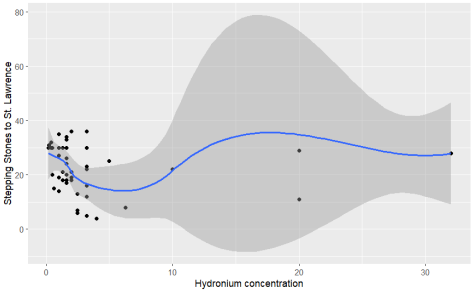

Figura 25: Diagramma di dispersione tra `Stepping stones to SL` e `Hydronium`, con curva LOESS in blu e intervallo di confidenza al 95% in grigio

IL diagramma di dispersione tra la distanza dei vari fiumi osservati dal fiume St. Lawrence e la concentrazione di nitrato nelle acque dei fiumi osservati *(Figura 24)* mostra una situazione capovolta rispetto a quella osservata nelle *Figura 16*, *Figura 20*, *Figura 23*. Il capovolgimento è dovuto al fatto che il fiume St. Lawrence è situato più a Nord-Est rispetto agli altri fiumi osservati. Quindi i valori che si trovavano distanti dagli altri fiumi, risultano vicini al fiume St. Lawrence e viceversa. La correlazione di Spearman è -0.41. La correlazione è negativa, ma al netto del segno, è in linea con le correlazioni tra gli altri fiumi e la concentrazione di nitrato. Il test per verificare se la correlazione è significativa (S=20062, p-value=0.005) porta a rifiutare l'ipotesi che la correlazione non sia significativa.

Dal diagramma di dispersione tra la distanza dei vari fiumi osservati dal fiume St. Lawrence e la concentrazione di idronio *(Figura 25)* si nota come la situazione, analogamente alla *Figura 24*, sia capovolta rispetto a quella osservata nelle *Figura 17* e *Figura 21*. La correlazione di Spearman è -0.35, indicando una lieve relazione negati-

va tra le due variabili. Il test per verificare se la correlazione è significativa (S=19135, p-value=0.02) porta a rifiutare l'ipotesi che la correlazione non sia significativa.

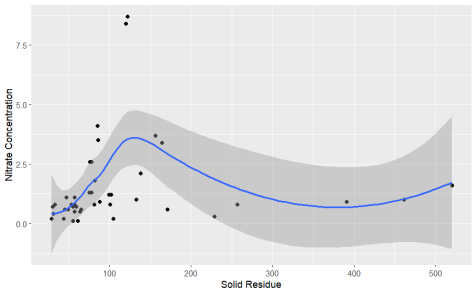

Figura 26: Diagrammma di dispersione tra `Nitrate` e `Solid Residue`, con curva LOESS in blu e intervallo di confidenza al 95% in grigio

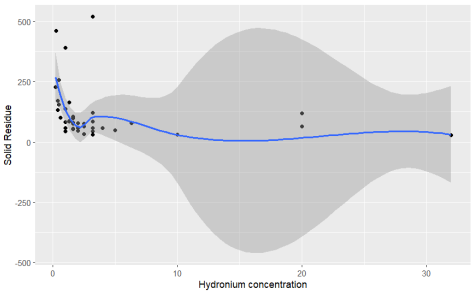

Figura 27: Diagramma di dispersione tra `Solid Residue` e `Hydronium`, con curva LOESS in blu e intervallo di confidenza al 95% in grigio

Dal diagramma di dispersione tra la concentrazione di nitrato e la concentrazione di residuo fisso nelle acque dei fiumi osservati *(Figura 26)* si osserva un iniziale aumento della concentrazione di idronio all'aumentare della concentrazione di residuo fisso. Tuttavia, dopo un certo punto (intorno al valore 150 del residuo fisso), la concentrazione di nitrato inizia a diminuire. La correlazione di Spearman è 0.50, indicando una modesta relazione positiva tra le due variabili. Il test per verificare se la correlazione è significativa (S=7386.3, p-value=0.001) porta a rifiutare l'ipotesi che la correlazione non sia significativa.

Dal diagramma di dispersione tra la concentrazione di residuo fisso e la variabile `Hydronium` *(Figura 27)* si nota una relazione molto ondulatoria tra le due variabili. La correlazione è nel complesso (tranne in una piccola parte del grafico tra i valori 2 e 3) negativa. La correlazione di Spearman è -0.55, confermando la presenza di una modesta relazione negativa tra le due variabili. Dunque, all'aumentare della concentrazione di residuo fisso diminuisce la concentrazione di idronio. Questo comportamento ha molto senso in quanto nel residuo fisso sono presenti alcuni sali minerali, come bicarbonato (HCO_3^-), carbonato (CO_3^{2-}) o cationi metallici (Ca_{2+}, *o* Mg_{2+}) che possono causare un effetto tampone neutralizzando gli ioni idrogeno, causando così una diminuzione della concentrazione di H^+ nell'acqua e a sua volta del pH. Il test per verificare se la correlazione è significativa (S=21968, p-value¡0.001) porta a rifiutare l'ipotesi che la correlazione non sia significativa.

2. Stima del modello

Per stimare il numero di specie di cozze trovate nei vari fiumi, si adattano diversi modelli di regressione.

2.1 Modello lineare multiplo

Come primo modello viene adattato un modello di regressione lineare normale multiplo*[7]*. Il modello è stato migliorato in fase di analisi, in particolare sono state rimosse in ordine le variabili `Stepping stones to AC` (p-value = 0.416), `Stepping stones to SL` (p-value = 0.394), `nitrate` (p-value = 0.171), `Stepping stones to SV` (p-value = 0.134).
È stata presa in considerazione la possibilità di includere un termine di interazione tra il numero di `Stepping stones to AP` e `area` (p-value = 0.01). Tuttavia, tale termine è stato escluso dal modello, in quanto aveva un coefficiente prossimo allo zero e, data la scarsa numerosità campionaria, non si è voluto appesantire il modello.

Il modello teorico è:

$$Sp = \beta_0 + \beta_1 \cdot A + \beta_2 \cdot AP + \beta_3 \cdot SR + \beta_4 \cdot H^+ + \epsilon_i \quad \text{con } \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

dove Sp è il numero di specie di cozze, AP è la distanza dal fiume Apalachicola, SR è la concentrazione di residuio fisso, H^+ è la concentrazione di ioni di idronio. Il vettore dei parametri ignoti di regressione è dato da $(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)^T$ mentre ϵ_i rappresenta il termine di errore.

Il modello stimato è riportato in *Tabella 3*.

	Stima	Std. Error	Intervallo di confidenza	t-value	p-value
β_0	15.451	1.889	(11.630, 19.271)	8.180	<0.001
β_1	0.0004	0.0001	(0.0002, 0.0006)	3.860	<0.001
β_2	-0.290	0.085	(-0.462, -0.118)	-3.415	0.002
β_3	-0.023	0.007	(-0.037, -0.010)	-3.497	0.001
β_4	-0.278	0.116	(-0.512, -0.044)	-2.408	0.021

Tabella 3: Adattamento del modello lineare multiplo

Tutte le stime dei parametri del modello risultano significative. Il modello stimato risulta:

$$\hat{Sp} = 15.451 + 0.0004 \cdot A - 0.290 \cdot AP - 0.023 \cdot SR - 0.278 \cdot H^+$$

Il coefficiente associato all'area è pari a 0.0004. Questo indica che, per ogni incremento unitario dell'area, si prevede un aumento medio di 0.0004 specie. La variabile `Stepping stones to AP` presenta un coefficiente stimato di -0.290, il che suggerisce che all'aumentare della distanza dal fiume Apalachicola il numero medio di specie diminuisce. La variabile `Solid Residue` ha un coefficiente pari a -0.023, indicando che un incremento unitario di `Solid Residue` comporta una riduzione media di

0.023 specie. Analogamente, la variabile `Hydronium` ha un coefficiente di -0.278, il che significa che, per ogni incremento unitario di `Hydronium`, il numero medio di specie diminuisce mediamente di 0.278 *(Tabella 3)*.

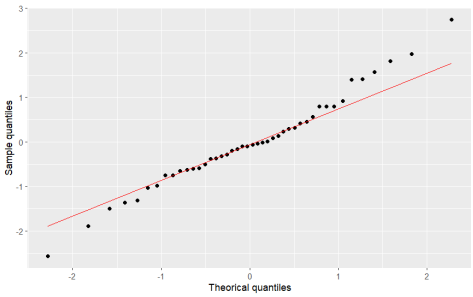

Figura 28: Diagramma quantile-quantile dei residui del modello

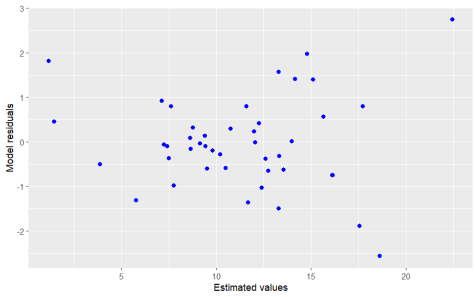

Figura 29: Diagramma di dispersione dei residui rispetto ai valori osservati

Sia il diagramma quantile-quantile *(Figura 28)*, sia il test di Shapiro-Wilk (W=0.984, p-value=0.777) confermano l'ipotesi di normalità dei residui del modello stimato. Il grafico dei residui studentizzati rispetto ai valori stimati *(Figura 29)* non mostra andamenti sistematici; pertanto, l'ipotesi di omoschedasticità non viene rifiutata. L'ipotesi di omogeneità, valutata attraverso la statistica test F, porta a un valore osservato della statistica test pari a 10.63 e un rispettivo p-value prossimo allo zero. Di conseguenza, l'ipotesi è stata rifiutata. L'indice di determinazione R^2 è pari a 0.522, indicando un moderato adattamento ai dati.

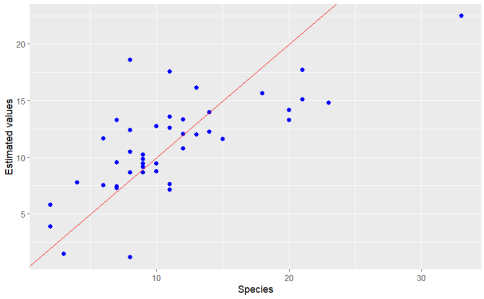

Figura 30: Diagramma di dispersione dei valori stimati rispetto ai valori osservati

Per valutare la bontà di adattamento del modello stimato, si confrontano i valori osservati della variabile specie con quelli predetti dal modello *(Figura 30)*. Il grafico mostra che i punti si distribuiscono principalmente lungo la bisettrice. Questo suggerisce un buon adattamento complessivo del modello ai dati. La correlazione tra i valori osservati e quelli predetti è pari a 0.722, indicando una forte relazione positiva. Questo risultato suggerisce che il modello riesce a spiegare una parte significativa della variabilità del numero di specie.

Nel corso dell'analisi, è stato valutato un modello alternativo in cui la variabile `Stepping stones to AC` è stata inclusa al posto di `Stepping stones to AP`. Questa scelta è stata motivata dal fatto che, una volta rimosso il termine `Stepping`

`stones to AP`, la variabile `Stepping stones to AC` risultava significativa. Tuttavia, tale modello è stato successivamente scartato in quanto presentava un valore di AIC maggiore (AIC = 261.12) rispetto al modello originale (AIC = 260.918), indicando una minore bontà di adattamento ai dati. Pertanto, si è preferito mantenere il modello iniziale.

2.2 Modello lineare generalizzato

Viene adattato un modello lineare generalizzato che assume che la variabile specie segua una distribuzione di Poisson con funzione di legame canonico*[8]*. In questo caso, il modello teorico assume che la media sia $\mu_i = e^{\eta_i}$ dove

$$\eta_i = \alpha_0 + \alpha_1 \cdot ln(Area) + \alpha_2 \cdot AP + \alpha_3 \cdot SR + \alpha_4 \cdot H^+,$$

e $(\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4)^T$ è il vettore dei parametri ignoti di regressione.

Anche in questo caso il modello è stato semplificato in fase di analisi; sono state eliminate le variabili `Stepping stones to SL` (p-value = 0.829), `Stepping stones to AC` (p-value = 0.891), `Stepping stones to SV` (p-value = 0.138), `nitrate` (p-value = 0.083). È stata considerata l'opportunità di includere un termine di interazione, questa volta tra la disteanza dal fiume Apalachicola (`Stepping stones to AP`) e la concentrazione di idronio (`Hydronium`) (p-value = 0.001). Tuttavia, si è deciso di non inserirlo, poiché l'analisi dei residui non ha fornito risultati soddisfacenti. Per il modello in esame, è stato adottato il logaritmo dell'area come variabile esplicativa, in quanto questa trasformazione ha portato a un miglioramento del modello. In particolare, il valore dell'AIC è diminuito da 245.83 a 241.54, indicando una migliore bontà di adattamento ai dati. Inoltre, la devianza residua è risultata ridotta, passando da 53.801 a 49.509, suggerendo una migliore capacità del modello di spiegare la variabilità del numero di specie.

Il modello stimato è riportato in *Tabella 4.*:

	Stima	Std. Error	Intervallo di confidenza	t-value	p-value
α_0	0.829	0.465	(-0.089, 1.734)	1.784	0.074
α_1	0.258	0.051	(0.159, 0.358)	5.098	<0.001
α_2	-0.025	0.006	(-0.036, -0.013)	-4.263	<0.001
α_3	-0.002	0.0006	(-0.003, -0.0009)	-3.376	<0.001
α_4	-0.032	0.011	(-0.055, -0.012)	-2.880	0.004

Tabella 4: Adattamento del modello di Poisson

Tutte le stime dei parametri del modello risultano significative, ad eccezione dell'intercetta.

Il modello stimato risulta:

$$\log(\hat{\mu}_i) = \eta_i = 0.829 + 0.258 \cdot \ln(Area) - 0.025 \cdot AP - 0.002 \cdot SR - 0.032 \cdot H^+$$

La variabile relativa al logaritmo dell'area ha un coefficiente stimato di 0.258. Questo implica che, per ogni incremento unitario del logaritmo dell'area, il valore atteso del logaritmo della risposta aumenta di 0.258, corrispondendo a un incremento

medio del numero di specie di circa $e^{0.258} = 1.294$ cioè del 29.4%. La distanza dal fiume Apalachicola ha un coefficiente pari a -0.0246, indicando che per ogni incremento unitario nella distanza dal fiume il valore atteso del logaritmo della risposta diminuisce di 0.0246, equivalente a una diminuzione media del numero di specie di circa il 2.4%. Per quanto riguarda il residuo fisso, il coefficiente stimato è -0.0021. Questo significa che ogni incremento unitario di residuo fisso determina una diminuzione del valore atteso del logaritmo della risposta di 0.0021, corrispondente a una riduzione media dello 0.2% nel numero di specie. Infine, la concentrazione di ioni idronio ha un coefficiente stimato di -0.0318. Questo implica che un incremento unitario di idronio riduce il valore atteso del logaritmo della risposta di 0.0318, con una diminuzione media del numero di specie di circa il 3.1% *(Tabella 4)*.

Il test contro il modello nullo porta un valore osservato della statistica test pari a 78.019 e un rispettivo p-value prossimo allo zero, confermando che il modello stimato è significativamente migliore rispetto a quello contenente solo l'intercetta. Inoltre, la devianza residua (49.509) risulta simile ai gradi di libertà (39), una ulteriore conferma dell'adeguatezza del modello.

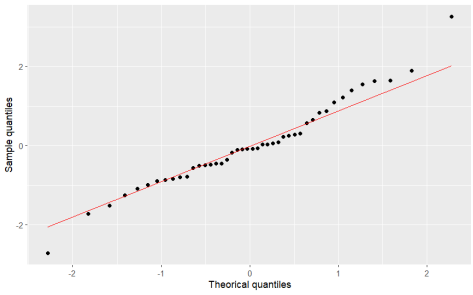

Figura 31: Diagramma quantile-quantile dei residui del modello

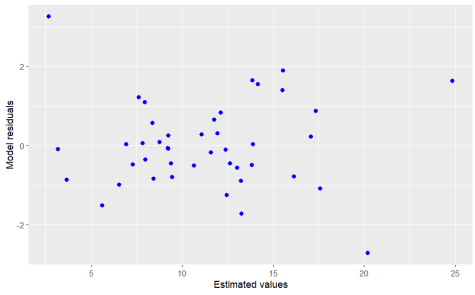

Figura 32: Diagramma di dispersione dei residui rispetto ai valori osservati

Sia il diagramma quantile-quantile *(Figura 31)*, sia il test di Shapiro-Wilk (W=0.972, p-value=0.361) confermano l'ipotesi di normalità dei residui del modello stimato. Il grafico dei residui rispetto ai valori stimati *(Figura 32)* non mostra andamenti sistematici; pertanto, l'ipotesi di omoschedasticità non viene rifiutata.

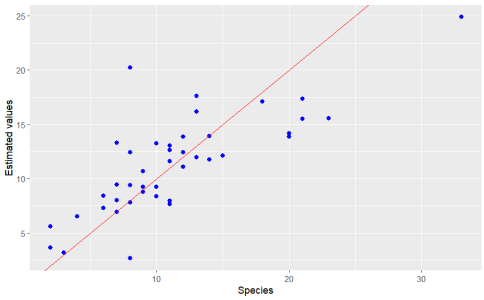

Figura 33: Diagramma di dispersione dei valori stimati rispetto ai valori osservati

Per valutare la bontà di adattamento del modello stimato, sono stati confrontati

i valori osservati della variabile `specie` con quelli predetti dal modello *(Figura 33)*. Il grafico evidenzia che i punti si distribuiscono principalmente lungo la bisettrice. Questo suggerisce un buon adattamento complessivo del modello ai dati. La correlazione tra i valori osservati e quelli predetti è pari a 0.779, indicando una forte relazione positiva. Questo suggerisce che il modello è in grado di spiegare una parte significativa della variabilità nel numero di specie.

Il test per la verifica della sovradispersione restituisce un valore osservato della statistica test pari a 50.568 e un p-value di 0.102; tale ipotesi viene quindi rifiutata.

Sono state valutate anche altre funzioni di legame, tra cui la radice quadrata ($\eta_i = \sqrt{\mu_i}$) e l'identità ($\eta_i = \mu_i$). Tuttavia, entrambe sono state scartate in quanto presentavano un valore di AIC superiore rispetto a quello del modello con funzione di legame canonica. In particolare, l'AIC del modello con legame canonico è 241.537, mentre quello con legame radice quadrata è 242.206 e quello con legame identità è 243.455.

3. Conclusioni

Il numero di specie di cozze è risultato significativamente influenzato dall'area di drenaggio del fiume, dalla distanza dal fiume Apalachicola, dalla quantità di minerali in soluzione e dalla concentrazione di ioni idronio.

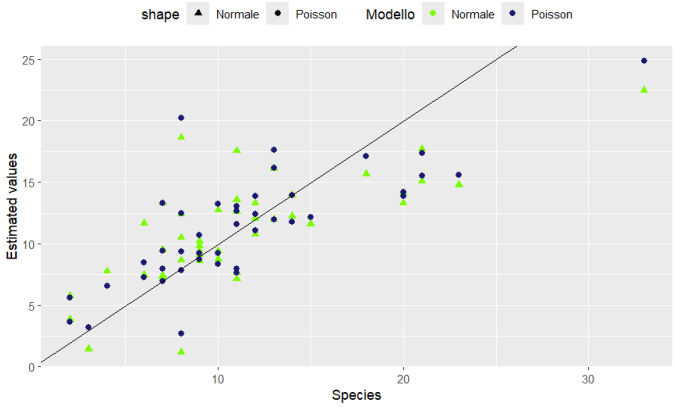

Figura 34: Diagramma di dispersione tra valori osservati e valori predetti con i due modelli

Entrambi i modelli stimati mostrano un comportamento complessivamente simile *(Figura 34)*, suggerendo che entrambi riescono a descrivere in modo adeguato le principali relazioni tra le variabili analizzate.

Confrontando gli AIC dei tre modelli, si osserva che il modello di Poisson presenta un valore di AIC inferiore (AIC = 241.537) rispetto al modello lineare normale

multiplo (AIC = 260.912). Di conseguenza, in base al criterio di Akaike, si preferisce il modello di Poisson.

Dal punto di vista pratico, i risultati suggeriscono che politiche di conservazione e gestione degli ecosistemi fluviali dovrebbero tenere conto di questi fattori. Ad esempio, la riduzione della concentrazione di ioni idronio e del residuo fisso potrebbe favorire un aumento della biodiversità delle specie di cozze. Inoltre, la vicinanza al fiume Apalachicola sembra svolgere un ruolo chiave nel sostenere un numero maggiore di specie, evidenziando l'importanza di preservare i diversi habitat fluviali.

Data l'importanza ecologica delle cozze di acqua dolce nel corso degli anni sono state oggetto di numerose ricerche. Esse hanno la capacità di filtrare l'acqua e alcune tossine svolgendo, fondamentalmente, un ruolo di depuratori naturali. Uno studio condotto nel 1993 sullo stato di conservazione delle cozze di acqua dolce in Stati Uniti e Canada*[9]* ha analizzato 297 specie native, anche a rischio di estinzione, e ha evidenziato alcuni fattori di minaccia per la sopravvivenza delle cozze di acqua dolce come la distruzione dell'habitat causato dalla costruzione di dighe o dal cambiamento dei canali d'acqua e dalla presenza di specie invasive. Un altro studio, più recente, pubblicato nel 2017 ha analizzato 16 specie di cozze di acqua dolce (2 specie di Margaritiferidae e 14 di Unionidae) in Europa*[10]* e ha evidenziiato ulteriori minacce come la presenza di inquinanti, la scomparsa di pesci ospiti e i pericoli legati al cambiamento climatico. Ulteriori sviluppi potrebbero approfondire il ruolo di questi fattori di rischio sulla diffusione delle specie di cozze di acqua dolce. Ci si potrebbe anche concentrare non più solo sul numero di specie, ma anche sulla numerosità della popolazione di ciascuna specie. O, ancora, dato che il dataset in analisi risale al 1974, uno studio potrebbe essere condotto aggiornando il dataset e valutando l'evoluzione delle specie di cozze di acqua dolce a 40 anni di distanza dallo studio originale.

4. Bibliografia

[1] Sepkoski J.J, Michael A. (1974). Distribution of Freshwater Mussels: Coastal Rivers as Biogeographic Islands, Systematic Biology, Volume 23, Issue 2, Pages 165-188. Rex.

[2] Britannica, T. Editors of Encyclopaedia (2025). Mussel. https://www.britannica.com/animal/mussel

[3] National Center for Biotechnology Information (2025). PubChem Compound Summary for CID 943, Nitrate. https://pubchem.ncbi.nlm.nih.gov/compound/Nitrate.

[4] Indiana Department of Natural Resources (2025). Freshwater Mussels. https://www.in.gov/dnr/fish-and-wildlife/wildlife-resources/animals/freshwater-mussels/

[5] Pfister C.A., Roy K., Wootton J.T., McCoy S.J., Paine R.T., Suchanek T.H., Sanford E. (2016) Historical baselines and the future of shell calcification for a foundation species in a changing ocean. Proc Biol Sci.

[6] National Center for Biotechnology Information (2025). PubChem Compound Summary for CID 123332, Oxonium. https://pubchem.ncbi.nlm.nih.gov/compound/Oxonium.

[7] Grigoletto M., Pauli F., Ventura L. (2017). Modello Lineare. Teoria e applicazione con R, Cap 4. Springer.

[8] Salvan A., Sartori N., Pace L. (2020). Modelli Lineari Generalizzati, Cap 5. Springer.

[9] Williams J., Jr Melvin, Cummings K., Harris J., Neves R. (1993). Conservation Status of Freshwater Mussels of The United States and Canada. Fisheries.

[10] Lopes-Lima M., Sousa R., Geist J., Aldridge D.C., Araujo R., Bergengren J., Bespalaya Y., Bódis E., Burlakova L., Van Damme D., Douda K., Froufe E., Georgiev D., Gumpinger C., Karatayev A., Kebapçi Ü., Killeen I., Lajtner J., Larsen B.M., Lauceri R., Legakis A., Lois S., Lundberg S., Moorkens E., Motte G., Nagel K.O., Ondina P., Outeiro A., Paunovic M., Prié V., von Proschwitz T., Riccardi N., Rudzīte M., Rudzītis M., Scheder C., Seddon M., Şereflişan H., Simić V., Sokolova S., Stoeckl K., Taskinen J., Teixeira A., Thielen F., Trichkova T., Varandas S., Vicentini H., Zajac K., Zajac T., Zogaris S. (2017) Conservation status of freshwater mussels in Europe: state of the art and future challenges. Biol Rev Camb Philos Soc.